Tijdschrift voor Economie en Management
Vol. L, 2, 2005

Inleiding tot robuuste statistiek: elementen van theorie en bedrijfseconomische toepassingen

door C. CROUX en K. JOOSSENS *

Christophe Croux
KULeuven; Departement Toegepaste Economische Wetenschappen, Vakgroep Kwantitatieve Methoden, Leuven.

Kristel Joossens
KULeuven; Departement Toegepaste Economische Wetenschappen, Vakgroep Kwantitatieve Methoden, Leuven.

ABSTRACT

Vele vaak gebruikte statistische methoden geven onbetrouwbare resultaten in de aanwezigheid van uitschieters. Robuuste statistische methoden blijven goed werken wanneer er atypische observaties aanwezig zijn of wanneer er niet perfect aan andere modelvoorwaarden voldaan is. Ofschoon de theorie van de robuuste statistiek zich reeds sinds enkele decennia ontwikkeld heeft, is het pas recentelijk dat robuuste schatters ook snel uitgerekend kunnen worden en in algemene statistische software pakketten opgenomen zijn. Toegepaste economen kunnen nu dan ook zonder problemen gebruik maken van robuuste schatters wanneer ze vrezen dat er uitschieters aanwezig zijn in hun gegevensbestanden. In dit artikel geven we een korte inleiding tot de theorie van de robuuste statistiek, aangevuld met verschillende voorbeelden uit de bedrijfseconomie.

* de auteurs willen Fentiane Haesbroeck van de Universiteit van Luik bedanken voor haar hulp bij de sectie met voorbeelden.

I. INLEIDING

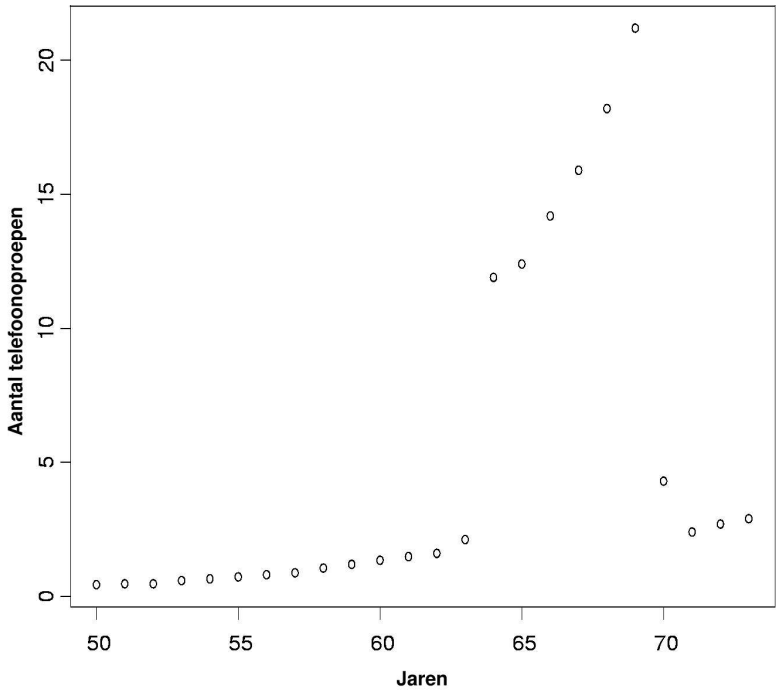
De wetenschap van de statistiek tracht bruikbare informatie te distilleren uit empirisch beschikbare gegevens. Om dit te realiseren, wendt men zich gedurende al meer dan twee eeuwen tot statistische modellen. Deze methode kende zijn apotheose in de eerste helft van de 20ste eeuw, vooral onder impuls van R.A. Fisher die een groot aantal statistische procedures introduceerde. Zijn werk vormt de basis van de inferentiële statistiek die men dagdagelijks gebruikt en die gebaseerd is op een parametrische specificatie van het statistische model.

eze klassieke benadering van de statistiek veronderstelt dat de statistische modellen goed gespecificeerd zijn. Sinds geruime tijd beseft men echter dat de reële wereld zich niet gedraagt zoals in de meeste vooropgestelde modellen. De performantie en validiteit van de toepassingen van parametrische procedures vereisen echter dat er strikt aan de hypotheses van het model voldaan is. Daarom werd de niet-parametrische statistiek geïntroduceerd en sommige van deze methodes zijn heel populair geworden in de toegepaste statistiek. Niettegenstaande het feit dat sommige problemen zeer bevredigend opgelost kunnen worden met een niet-parametrische methode, heeft de parametrische aanpak nog steeds een dominante rol omdat zij vaak meer precies is en de geschatte parameters vaak een (economische) interpretatie hebben. Bovendien zijn parametrische procedures in een veel groter gamma van situaties toepasbaar.

De robuuste statistiek combineert de kracht van beide benaderingen. Zij doet niet alleen dienst in parametrische modellen, maar zij gebruikt ook procedures die minder essentieel steunen op de hypotheses waaraan het gekozen model moet voldoen. Bovendien laten robuuste methodes toe om afwijkende observaties te identificeren. De robuuste statistiek gaat ervan uit dat de meest voorkomende hypotheses in de statistiek (zoals normaliteit, lineariteit, ...) enkel bij benadering juist zijn. Haar doel is dus het creëren van procedures die weerstand bieden aan zulke modelafwijkingen.

Laten we een voorbeeld geven van een dataset waarin uitschieters (*outliers*) voorkomen. Voor de periode 1950-1973 werd jaarlijks de duurtijd in minuten van internationale telefoonoproepen in België gemeten (zie Rousseeuw en Leroy (1987)). Deze gegevens worden voorgesteld in Figuur 1. Deze tijdsreeks bevat een aantal zeer atypische observaties van 1964 tot 1969, wat te wijten is aan het feite een ander registratiesysteem gebruikt werd in die periode. In die periode werd immers het aantal telefoongesprekken gemeten en niet de totale duurtijd. In dit voorbeeld komen er dus “grove fouten” voor, die uitschieters genereren. Uitschieters zijn observaties die zich anders gedragen dan de grote meerderheid van de andere gegevens en waarvan het zeer onwaarschijnlijk is dat ze door hetzelfde proces gegenereerd zijn als de grote meerderheid van de andere observaties. In een robuuste procedure gaat men deze uitschieters dan ook een kleiner gewicht geven, of soms zelfs helemaal weglaten, zodat zij weinig of zelfs geen invloed hebben op de analyse.

FIGUUR 1
Aantal telefoonoproepen van 1950 tot 1973 in België



In dit voorbeeld kan men de uitschieters gemakkelijk grafisch detecteren, ze bevinden zich immers ver van de meerderheid van de gegevens in de grafiek van Figuur 1. Het detecteren van uitschieters is niet altijd zo eenvoudig. Indien we werken met multivariate gegevens, bijvoorbeeld observaties in 5 dimensies, dan wordt het onmogelijk de data grafisch voor te stellen en kunnen uitschieters niet meer visueel gedetecteerd worden. Daarom is nuttig om robuuste procedures te gebruiken en om detectieprocedures voor uitschieters te ontwikkelen. Merk op dat de uitschieters soms juist de interessantste observaties zijn, omdat ze met speciale gebeurtenissen overeenkomen.

Het probleem van robuustheid is reeds lang gekend en statistici zijn zich sterk bewust van de gevaren van uitschieters. Op de middelbare school leren scholieren reeds dat de mediaan meer bestand is tegen uitschieters dan het gemiddelde. Toch heeft het relatief lang geduurd vooraleer men een meer formele benadering van het probleem vond. Pionierswerk van Huber (1964) en Hampel (1971) introduceerde maten om de robuustheid van een schatter te meten. Sindsdien zijn er tal van theoretische ontwikkelingen gebeurd en

werden vele nieuwe technieken geïntroduceerd die resistent zijn tegen uitschieters.

Dit artikel geeft een eerste kennismaking met robuuste statistiek. In Sectie II komen de basisbegrippen invloedsfunctie en breekpunt van een schatter aan bod. Dit zijn maten van robuustheid die toelaten om de robuustheid van een schatter te evalueren. In deze sectie beperken we ons tot ééndimensionale gegevens, om zodoende de uiteenzetting eenvoudig te kunnen houden. In Sectie III wordt het lineaire regressiemodel behandeld. We zullen aantonen dat de klassieke kleinste kwadraten schatter niet robuust is, en een alternatieve schattingmethode bespreken. In Sectie IV illustreren we de voordelen van een robuuste aanpak met enkele bedrijfseconomische toepassingen. In een laatste sectie maken we de nodige verdere verwijzingen naar de wetenschappelijke literatuur in dit onderzoeksdomein.

Basisbegrippen van robuuste statistiek

Beschouwen we een steekproef van univariate gegevens, die we noteren als $X=\{x_1,\ldots,x_n\}$. Onderstel dat de populatie waaruit deze gegevens getrokken worden normaal verdeeld is met gemiddelde μ en een variantie σ^2. We willen nu de parameter μ, die hier de centrale waarde van de verdeling aangeeft, schatten. De klassieke schatter voor μ is het rekenkundig gemiddelde, gedefinieerd als

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

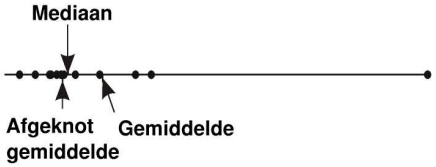
Deze schatter wordt echter sterk beïnvloed door één of meerdere uitschieters in onze steekproef. Beschouw volgende reeks van bruto maandinkomens van 12 werknemers van een zeker bedrijf (in euro).

$X=\{1513, 1834, 2112, 2160, 2288, 2375, 2424, 2647, 3156, 3908, 4233, 9961\}.$

Het rekenkundig gemiddelde van deze steekproef van inkomens bedraagt $\bar{x} = 3217.58$ euro per maand, wat duidelijk afwijkt van het centrum van de data, zoals gemeten door de mediaan. Dit wordt geïllustreerd in Figuur 2. Het gemiddelde werd hier sterk aangetast door de atypische extreme waarde, 9961, en is hier dus geen goede schatter voor de parameter μ van de vooropgestelde normale verdeling. Een meer robuuste schatter is nodig[1].

FIGUUR 2

Inkomensreeks van 12 werknemers van een firma



Indien we in een gegeven steekproef uitschieters verwachten, kunnen we verschillende strategieën toepassen. Het rekenkundig gemiddelde is immers niet de enig mogelijke schatter voor μ, vele alternatieven bestaan. Een van de meest gebruikte is zonder twijfel de mediaan, die gelijk is aan de "middelste" observatie in de steekproef. Een formele definitie wordt gegeven door

$$\text{med}_{\text{i}}\ x_i = \frac{1}{2}(x_{(\lfloor n/2 \rfloor+1)} + x_{(\lfloor (n+1)/2 \rfloor)}),$$

waar $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ de geordende observaties zijn en $\lfloor z \rfloor$ correspondeert met het grootste geheel getal kleiner of gelijk aan z. Een andere mogelijke schatter is het afgeknot gemiddelde: voor een reële waarde α tussen 0 en 0.5, wordt het afgeknot gemiddelde met drempel α gedefinieerd als het rekenkundig gemiddelde berekend op basis van een "afgeknotte" steekproef. De afgeknotte steekproef is de steekproef waaruit we de $\lfloor \alpha n \rfloor$ kleinste en de $\lfloor \alpha n \rfloor$ grootste observaties laten wegvallen. We noteren dit afgeknot gemiddelde met $\bar{x}_\alpha$ en een wiskundige definitie is

$$\bar{x}_\alpha = \frac{1}{n - 2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor+1}^{n-\lfloor \alpha n \rfloor} x_{(i)}.$$

De waarde α dient door de statisticus zelf gekozen te worden. Als verwacht wordt dat er in de dataset veel atypische observaties kunnen voorkomen, is het aangewezen om α groot te nemen. Hoe groter de waarde van α, hoe meer efficiëntie men echter verliest. Een goede keuze voor α wordt gegeven door α = 0.25 (cfr. Croux en Haesbroeck (2002)). Een afgeknot gemiddelde met drempelwaarde 25% resulteert in een schatter met een grote robuustheid, en tegelijk een precisie die bijna zo groot is als die van het steekproefgemiddelde (in afwezigheid van uitschieters).

Merk op dat een afgeknot gemiddelde met drempel $\alpha \approx 0.5$ overeen komt met de mediaan. In het voorbeeld van Figuur is de mediaan van de inkomens 2399.5 terwijl het afgeknot gemiddelde met drempel 25% gegeven wordt door $\bar{x}_{0.25}$ = 2508.33.

Omdat er vele alternatieven zijn voor het rekenkundig gemiddelde, is het belangrijk dat we hun performanties kunnen vergelijken aan de hand van verschillende criteria. Vaak wordt als criterium de efficiëntie van de schatter genomen. Hoe efficiënter een schatter, hoe preciezer hij de onbekende μ zal schatten. Men kan aantonen dat het rekenkundig gemiddelde, onder de assumptie van normaliteit, de meest efficiënte schatter is. Het is echter zo dat het gemiddelde deze eigenschap snel verliest en helemaal niet meer zo precies is wanneer er afwijkingen van het model zijn. Daarom is het ook nodig om andere maten van performantie van een schatter te bekijken. In de volgende twee paragrafen worden twee manieren voorgesteld om de robuustheid van een schatter te meten.

A. *De empirische invloedsfunctie*

De empirische invloedsfunctie (EIF) laat toe het effect van een afwijkende observatie op de schatter te visualiseren. Gegeven een steekproef $x_1,\ldots,x_n$ en een schatter T. Voor elke mogelijke waarde van x berekenen we dan

$$\mathrm{EIF}(x;T) = n\ \{T(x_1,\ldots,x_n, x)\tilde{}\ T(x_1,x_2,\ldots,x_n)\}.$$

(In bovenstaande formule is de vermenigvuldiging met de steekproefgrootte n enkel een herschaling). De empirische invloedsfunctie laat toe het effect op de schatter T te meten, wanneer een observatie x aan de steekproef wordt toegevoegd. Wanneer de schatter robuust is, zou dit effect beperkt moeten blijven. We willen immers niet dat individuele observaties, die mogelijke uitschieters kunnen zijn, teveel invloed op onze schatter uitoefenen.

Voor het voorbeeld met de inkomens, waar we als steekproef de eerste 11 observaties zonder de uitschieter nemen, hebben we empirische functies uitgerekend voor het rekenkundig gemiddelde $\bar{x}$, de mediaan, en het 25% afgeknot gemiddelde $\bar{x}_{0.25}$ (Figuur 2). We stellen onmiddellijk vast dat de EIF van het rekenkundig gemiddelde onbegrensd is. Grote waarden hebben een onbegrensde invloed op $\bar{x}$, wat de niet-robuustheid van het gemiddelde aantoont. Noem

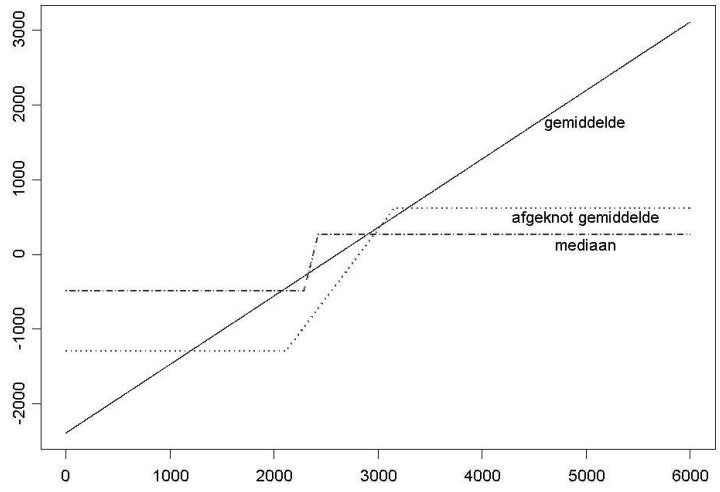
$$\gamma(T) = \sup_x |\ \mathrm{EIF}(x;T)\ | \qquad (1)$$

de maximale waarde die de EIF kan aannemen. Dit getal wordt ook de *gross-error sensitiviteit* genoemd, en is een maat voor de robuustheid van een schatter. Hoe kleiner deze waarde is, hoe beter. Uit Figuur 3 blijkt dat de empirische invloedsfunctie van de mediaan en het afgeknot gemiddelde begrensd zijn. De waarden ervan zijn respectievelijk $\gamma(med) = 478.5$ en $\gamma(\bar{x}_{0.25}) = 1291.19$, terwijl $\gamma(\bar{x}) = \infty$.

Indien men enkel de gross-error sensitiviteit als een maat voor robuustheid neemt, is de mediaan te verkiezen boven het afgeknot gemiddelde

met drempel 25%. Het rekenkundig gemiddelde heeft geen begrensde invloedsfunctie, wat zijn niet-robuustheid aantoont.

FIGUUR 3
Empirische invloedsfuncties voor het rekenkundig gemiddelde (volle lijn), de mediaan (gestreepte lijn) en het afgeknotte gemiddelde met drempel 25% (stippellijn)



De empirische invloedsfunctie en bijhorende gross-error sensitiviteit zijn gemakkelijk te berekenen, maar een nadeel van formule (1) is dat deze waarde nog afhangt van de gegevens $x_1,\ldots,x_n$. Met behulp van statistische functionalen en verdelingen, kan men theoretisch werkbaardere definities van invloedsfuncties en gross-error sensitiviteit introduceren (Hampel et al (1986)). Het achterliggend idee is echter hetzelfde als hierboven beschreven. De empirische invloedsfunctie meet de gevoeligheid van een schatter voor individuele observaties. Het blijkt nu dat in datasets vaak meerdere uitschieters tegelijk voorkomen; men spreekt dan van clusters van atypische observaties. Een meer geschikte maat van robuustheid in dit kader is dan het breekpunt, dat in de volgende paragraaf gedefinieerd wordt.

B. *Breekpunt*

Het breekpunt van een schatter T is de kleinste fractie observaties die we moeten wijzigen opdat de schatter willekeurig grote waarden kan aannemen. Om het breekpunt van een schatter T te vinden voor een steekproef $X=\{x_1,\ldots,x_n\}$, gaat men als volgt te werk. Vertrekkende van de initiële

steekproef X creëren we een gecontamineerde steekproef X' door m observaties van X te veranderen in willekeurig (grote) waarden. Dit creëert dan een vertekening of *bias* die gegeven wordt door $|T(X)-T(X')|$.

Bedoeling is nu om de m observaties dusdanig te veranderen zodat deze bias zo groot mogelijk wordt. Deze maximale bias van de schatter T die men kan verkrijgen door m observaties te wijzigen is dan

$$\text{maxbias}\,(m,T,X) = \sup_{X'} |T(X)-T(X')|. \qquad (2)$$

Als deze maxbias oneindig groot is, zegt men dat de schatter "breekt", hij neemt een volstrekt onbetrouwbare waarde aan. Het breekpunt $\varepsilon^*(T)$ van de schatter T is nu de kleinste fractie m/n van observaties die men moet veranderen alvorens de bias oneindig groot wordt, en een wiskundige definitie is

$$\varepsilon^*(T) = \frac{1}{n} \min \{m : \text{maxbias}\,(m,T,X) = \infty\}.$$

Het is nu niet moeilijk om de breekpunten van de beschouwde schatters te berekenen. Kijken we naar Figuur 1 en beelden we ons in dat we één enkele observatie naar oneindig verplaatsen. Dan zal het rekenkundig gemiddelde ook mee oneindig groot worden, en we krijgen dus $\varepsilon^*(\bar{X}) = 1/12$. Het verplaatsen van deze observatie naar oneindig gaat echter de mediaan en het afgeknot gemiddelde niet laten breken. Voor de mediaan moeten we maar liefst 6 observaties naar oneindig laten gaan, terwijl het voor een 25% afgeknot gemiddelde slechts 4 observaties gecontamineerd moeten worden om deze schatter te laten breken en dus een oneindig grote waarde te laten aannemen.

In ons voorbeeld hadden we een kleine steekproef. Het is echter niet moeilijk om in te zien dat voor zeer grote steekproeven geldt $\varepsilon^*(\bar{X}) \approx 0$, $\varepsilon^*(med) \approx 0.5$ en $\varepsilon^*(\bar{X}_{0.25}) \approx 0.25$. Indien we het breekpunt van een schatter als maat voor robuustheid nemen, is de mediaan weer te verkiezen is boven het 25% afgeknot gemiddelde. Het gewone gemiddelde heeft een breekpunt gelijk aan nul, wat nogmaals de niet-robuustheid van deze schatter aantoont.

Merken we tot slot nog op dat definitie (2) afhangt van de gekozen steekproef. Om een theoretisch meer werkbare definitie van de maximale bias te krijgen, zal het weer nodig zijn om met verdelingstheorie en statistische functionalen te werken. We gaan hier niet verder op in. In deze sectie werden verschillende maten voor robuustheid besproken. Bij de keuze van een geschikte schatter zal men echter niet enkel zijn robuustheid beschouwen maar ook zijn precisie en berekenbaarheid.

III. ROBUUSTE LINEAIRE REGRESSIE

Het afgeknot gemiddelde en dus ook de mediaan, die we in de vorige sectie bespraken, zijn goed gekende robuuste schatters om de centrale positie van een (symmetrische) univariate verdeling te schatten. Het is echter minder duidelijk hoe men een robuuste schatter kan bekomen voor meer complexe modellen, zoals het lineaire regressiemodel. Voor een steekproefgrootte n meten we hier waarden y_i van een te verklaren variabele, en waarden $x_{i1},\ldots,x_{ip}$ van de verklarende variabelen voor elke observatie $i=1,\ldots,n$. Er wordt verondersteld dat de relatie tussen de te verklaren variabelen en de verklarende variabelen lineair is, dus

$$y_i=\beta+x_{i1}\beta_1+\ldots+x_{ip}\beta_p+e_i \qquad \text{voor } i=1,\ldots,n. \qquad (3)$$

De storingtermen $e_1,\ldots,e_n$ worden verondersteld om onafhankelijk en identiek verdeeld te zijn. Vaak wordt daarboven de hypothese van normaliteit voor deze storingstermen opgelegd. De onbekende parameters in het regressiemodel zijn de constante term β, en de richtingscoëfficiënten $\beta_1,\ldots,\beta_p$. We noteren nu de vector van ongekende parameters als

$$\theta = (\beta,\beta_1,\beta_2,\ldots,\beta_p)$$

en de bedoeling is om deze ongekende parametervector te schatten met behulp van de beschikbare data. Het residu van de i-de observatie wordt gegeven door

$$r_i(\theta) = y_i-(\alpha+x_{i1}\beta_1+\ldots+x_{ip}\beta_p).$$

Bedoeling is nu om een schatter $\hat{\theta}$ zo te kiezen dat deze residuen zo klein mogelijk zijn. Met zo "klein" mogelijk, wordt bedoeld dat voor een gekozen doelfunctie f, de waarde van de doelfunctie uitgerekend in de residuen minimaal wordt, m.a.w.

$$\hat{\theta} = \arg\min_{\theta} f(r_1(\theta),\ldots,r_n(\theta)). \qquad (4)$$

Als doelfunctie wordt meestal de som van de gekwadrateerde residuen genomen, wat resulteert in

$$\hat{\theta}_{\mathrm{LS}} = \arg\min_{\theta} \sum_{i=1}^{n} r_i^2(\theta). \qquad (5)$$

Dit geeft dan de bekende methode van de kleinste kwadraten, of *Least Squares* (LS). Het is echter belangrijk te weten dat dit niet de enige mogelijke

schatter voor het lineaire regressiemodel is. Andere doelfuncties f zullen resulteren in andere schatters.

De reden van de populariteit van de kleinste kwadraten schatter is historisch te verklaren: toen men rond 1800 lineaire modellen begon te beschouwen, was de kleinste kwadraten schatter de enige die men vrij eenvoudig kon uitrekenen. Gauss schreef: "Van alle principes is de kleinste kwadraten het eenvoudigste: voor de anderen moeten we complexe berekeningen maken." Daarna introduceerde Gauss de normale verdeling als zijnde de verdeling waarvoor de kleinste kwadraten schatter optimaal is, in de zin van maximale efficiëntie. Sindsdien is de combinatie van de hypothese van normaliteit en het gebruik van de kleinste kwadraten schatter standaard. Door de beschikbaarheid van computers is het nu echter mogelijk geworden om (4) ook te berekenen voor andere doelfuncties *f*. Bovendien hebben statistici zich gerealiseerd dat gegevens vaak niet aan de klassieke normaliteitshypothese voldoen, en dat optimaliteit dus niet gegarandeerd is. In het bijzonder is het geweten dat de kleinste kwadraten methode zeer kwetsbaar is voor uitschieters. In de volgende paragraaf besteden we aandacht aan de verschillende soorten uitschieters die in een regressieanalyse kunnen optreden.

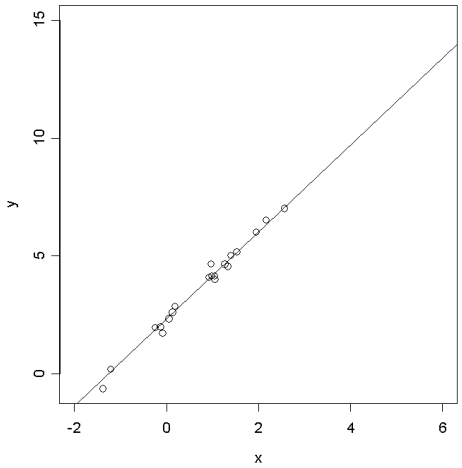
A. *Verticale uitschieters en hefboompunten*

In de context van regressie komen twee soorten uitschieters voor, namelijk verticale uitschieters en hefboompunten. Als illustratie, beschouw een eenvoudig regressiemodel $y_i = \alpha + \beta x_i + e_i$ met slechts één verklarende variabele. Een fictieve dataset, die in Figuur 4 (a) wordt voorgesteld, werd gegenereerd volgens dit model. De gegevens kunnen hier in het vlak worden voorgesteld en we zien dat er geen uitschieters aanwezig zijn. Na schatting met de kleinste kwadraten methode bekomen we de regressierechte $y = \hat{\alpha} + \hat{\beta} x$, en in Figuur 4 (a) zien we dat deze een goede fit geeft voor de puntenwolk.

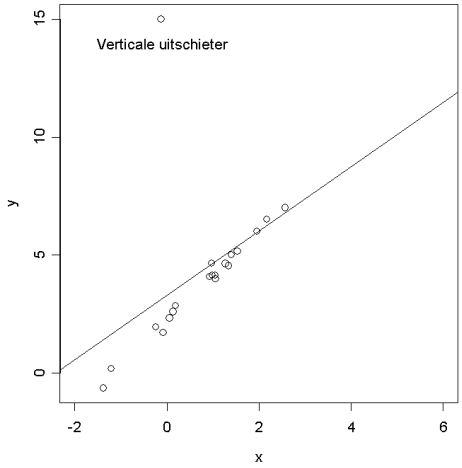
Indien we nu één van de observaties in verticale richting verschuiven, dan krijgen we een *verticale uitschieter*, zoals in Figuur 4 (b). We merken onmiddellijk op dat de geschatte regressierechte nu een veel minder goede fit geeft. Het toont reeds aan dat kleinste kwadraten schatter door slechts één enkele uitschieter sterk kan veranderen.

FIGUUR 4
Effect van uitschieters op de kleinste kwadraten schatter:
(a) geen uitschieters (b) verticale uitschieter
(c) slecht hefboompunt (d) goed hefboompunt

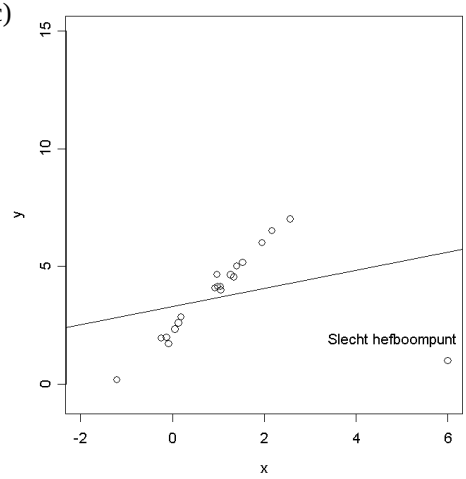
(a)



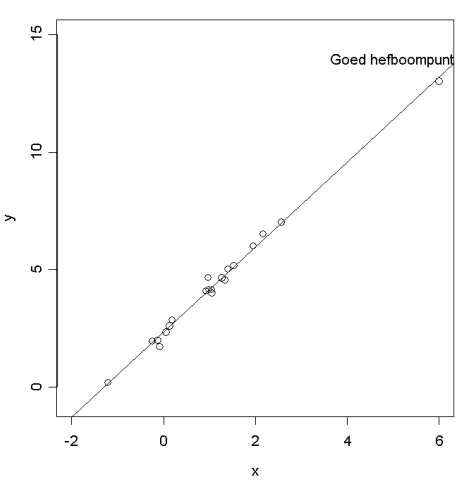
(b)



(c)



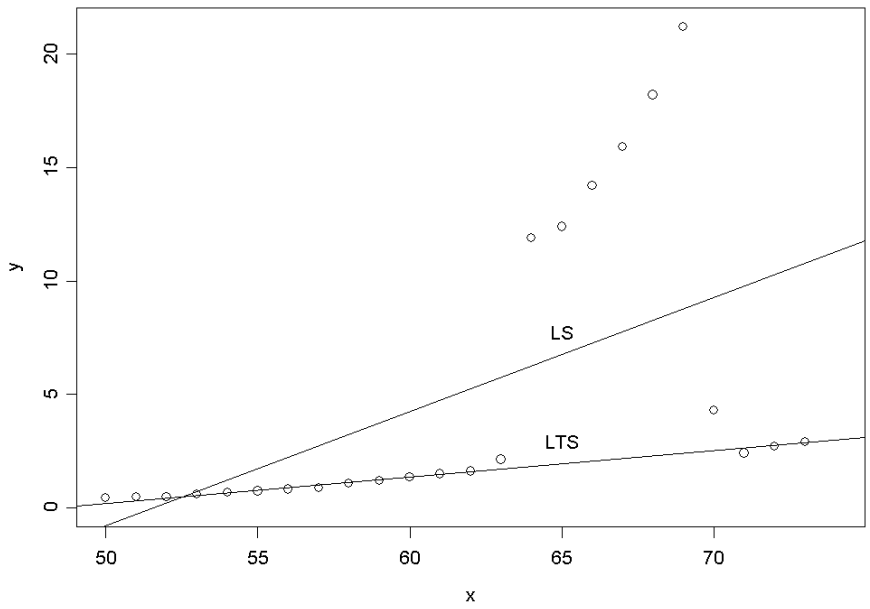
d)



Wanneer de uitschieter zodanig is dat de waarde van x_i atypisch is in de ruimte van de verklarende variabelen, dan spreekt men van een *hefboompunt*. In Figuur 4 (c) en Figuur 4 (d) zien we zulk een hefboompunt: de corresponderende *x*-waarde is inderdaad ver weg van de grote meerderheid van andere punten op de *x*-as. In Figuur 4 (c) zien we dat het hefboompunt erin slaagt om de geschatte regressierechte naar zich toe te trekken: de rechte kantelt zoals een hefboom. In Figuur 4 (d) heeft het hefboompunt schijnbaar zo goed als geen effect op de kleinste kwadraten schatting. Dit komt omdat de uitschieter nog steeds de lineaire relatie volgt die de andere punten ook volgen, en de regressierechte dus niet doet kantelen. We noemen dit een *goed hefboompunt*, terwijl we in Figuur 4 (c) spreken van een *slecht hefboompunt*. Slechte hefboompunten zijn het meest gevaarlijk, en oefenen meer invloed uit dan verticale uitschieters.

Keren we nu terug naar ons eerste voorbeeld van de telefoondata. In Figuur 5 zien we de gegevens met de geschatte kleinste kwadraten regressierechte. Er is duidelijk aanwezigheid van verticale uitschieters, en de LS schatter wordt hierdoor sterk vertekend. We kunnen nog moeilijk zeggen dat we een goede fit bekomen voor de data. Merk ook op dat meerdere residuen voor de goede observaties groter zijn dan de residuen voor de verticale uitschieters. Dit betekent dat gebruik maken van de grootte van de residuen -berekend ten opzichte van de regressierechte - om uitschieters te detecteren geen goede techniek is. Residuele analyse kan erg misleidend zijn: uitschieters kunnen kleine residuen hebben (dit noemt men *masking*) en goede observaties grote residuen (dit noemt men *swamping*). Wanneer de residuen echter berekend worden ten opzichte van een robuust geschatte regressierechte, kan men de grootte van de residuen wel gebruiken voor detectie van uitschieters. In Figuur 5, waar we ook een robuust geschatte regressierechte getekend hebben, zien we onmiddellijk dat de uitschieters veel grotere residuen hebben dan de goede observaties.

Figuur 5
Regressierechte voor de telefoondata geschat met de kleinste kwadraten methode (LS) en met een robuuste schatter (LTS)



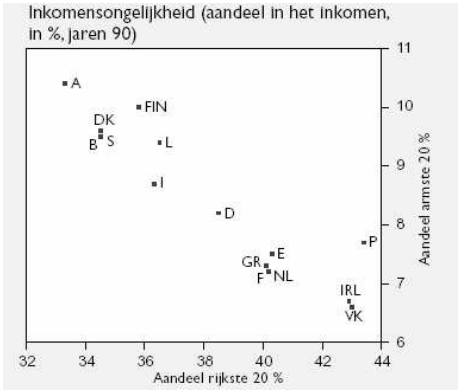
Uit het bovenstaande kunnen we besluiten dat één uitschieter voldoende is om de kleinste kwadraten schatter zeer sterk te beïnvloeden. Men kan aantonen dat LS een onbegrensde invloedsfunctie en een breekpunt van nul heeft, net zoals het rekenkundig gemiddeld.

Hefboompunten en verticale uitschieters kunnen ook voorkomen in economische data. KBC Bank en Verzekeringen (2001) bestudeerde het "Economisch profiel van de Europese Unie", en we vinden in hun rapport verschillende dispersiediagrammen (of *scatterplots*) terug. Een selectie hiervan presenteren we in Figuur 6.

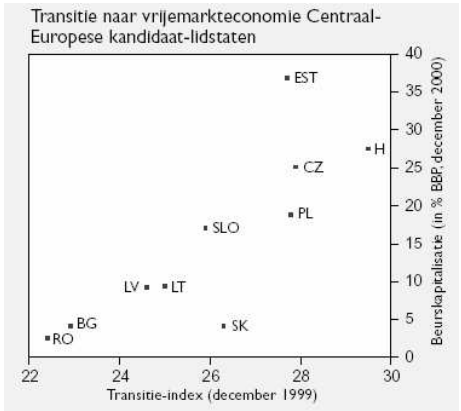
FIGUUR 6

Voorbeelden van dispersiediagrammen met verschillende types uitschieters

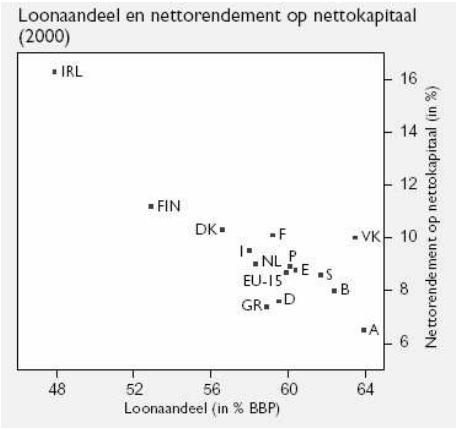
(a)



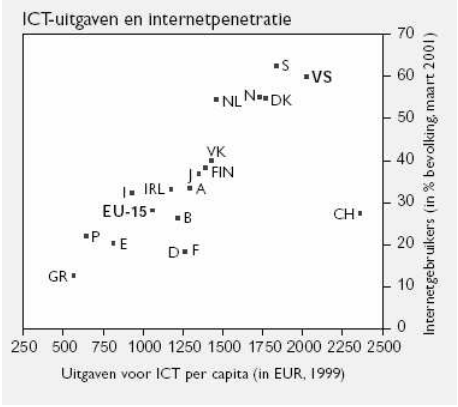
(b)



(c)



d)



(*Bron*: KBC-studiedienst)

In Figuur 6 (a) zien we een duidelijk dalende relatie, en geen opvallende uitschieters. Figuur 6 (b) toont een stijgende relatie, maar Estland (EST) is hier een verticale uitschieter. Dit diagram stelt Oost-Europese landen voor, en de beurskapitalisatie in Estland is veel groter dan men op basis van zijn transitie index mag verwachten. In Figuur 6 (c) zien we een voorbeeld van een goed hefboompunt: Ierland (IRL) volgt de lineaire relatie die de andere landen ook volgen, maar heeft een extreem lage waarde voor de verklarende variabele loonaandeel. Tot slot zien we in Figuur 6 (d) dat Zwitserland (CH) hier een slecht hefboompunt is. Het heeft de grootste waarde voor de *x*-variabele, maar volgt de lineaire relatie tussen "uitgaven voor ICT per capita" en "internetgebruikers" niet. In meerdere empirische studies blijkt het dat landen als Zwitserland en Luxemburg vaak als uitschieter gedetecteerd worden. Ze gedragen zich anders dan de meerderheid van andere Europese landen, en de statistische analyse mag hierdoor niet teveel beïnvloed worden.

B. *Robuuste schatters*

Door een geschikte doelfunctie *f* te kiezen in (4) is het mogelijk om robuuste schatters te bekomen voor het lineaire regressiemodel. Een reden waarom de kleinste kwadraten schatter erg beïnvloed wordt door uitschieters is dat het *kwadraat* van de residuen in de doelfunctie optreedt, waardoor hun effect nog vergroot wordt. In plaats van het kwadraat kan men ook de absolute waarden van de residuen opnemen in de doelfunctie. Zo bekomt men de kleinste absolute waarde of *Least Absolute Value* schatter

$$\hat{\theta}_{\mathrm{LAV}} = \arg\min_{\theta} \sum_{i=1}^{n} \left| r_i(\theta) \right| . \qquad (6)$$

In tegenstelling tot de kleinste kwadraten methode biedt $\hat{\theta}_{\mathrm{LAV}}$ bescherming tegen de aanwezigheid van verticale uitschieters, maar blijft gevoelig voor slechte hefboompunten. Men kan aantonen dat het breekpunt van deze schatter ook 0% is.

Om een werkelijk robuuste methode met een hoog breekpunt te bekomen moet men in (5) de som door een mediaan vervangen. De bekomen schatter is dan de *Least Median of Squares* regressie schatter van Rousseeuw (1984)

$$\hat{\theta}_{\mathrm{LMS}} = \arg\min_{\theta} \operatorname{med}_{i} r_i^2(\theta) . \qquad (7)$$

Een ander voorstel bestaat erin om in de doelfunctie van de kleinste kwadraten schatter niet de som over alle residuen in het kwadraat te nemen, maar enkel de som over de kleinste *h*. Men kiest dan $h = \lfloor n(1+\alpha) \rfloor$, met α een drempel waarde tussen 0 en 0.5. De doelfunctie is dan een afgeknotte som van residuen in het kwadraat, en men bekomt de *Least Trimmed Squares* (LTS) schatter

$$\hat{\theta}_{\mathrm{LTS}} = \arg\min_{\theta} \sum_{i=1}^{h} r_{(i)}^2(\theta) , \qquad (8)$$

met $r_{(1)}^2 \leq \ldots \leq r_{(n)}^2$. Als keuze voor α kan men bijvoorbeeld $\alpha = 0.25$ nemen, wat betekent dat men het grootste kwart van de gekwadrateerde residuen niet laat meespelen in de doelfunctie. Een andere mogelijke keuze is $\alpha = 0.50$, wat leidt tot het hoogst mogelijke breekpunt van 50%.

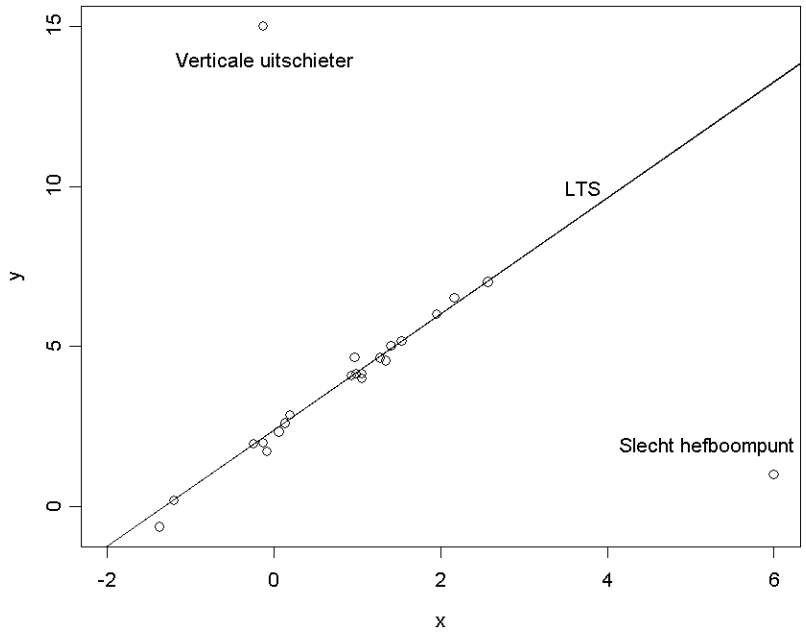
Ofschoon de definitie van de LMS en LTS schatter vrij eenvoudig is, zijn beide moeilijk uit te rekenen. Sinds enkele jaren zijn echter snelle

algoritmes beschikbaar die deze schatters kunnen uitrekenen en die geïmplementeerd werden in statistische software. Er is een voorkeur voor de LTS schatter omdat deze statistisch efficiënter is en sneller te berekenen. De LMS heeft echter een kleinere maximale bias en is in die zin robuuster dan de LTS.

Om de robuustheid van deze schatters te illustreren keren we terug naar de telefoondata (Figuur 5), waar ook de rechte bekomen door de LTS schatter (met $\alpha = 0.5$) weergegeven is. We zien dat de robuuste methode de lineaire relatie, die de grote meerderheid van de observaties volgt, terugvindt. Grote residuen ten op zichte van deze regressierechte geven ons dan de uitschieters. Eens deze uitschieters gedetecteerd zijn, kan men trachten op te sporen waarom deze observaties zich vreemd gedragen.

Ook op de artificiële data van Figuur 4 kunnen we een robuuste schatter toepassen. We geven hier 3 configuraties van de gegevens (die zonder uitschieters, die met een verticale uitschieter en die met een slecht hefboompunt), samen met drie door LTS (met $\alpha = 0.5$) geschatte regressierechten op één enkele tekening in Figuur 7. De drie regressierechten voor deze 3 configuraties zijn praktisch niet verschillend, waardoor ze op de tekening niet te onderscheiden zijn. De uitschieters hebben dus nauwelijks effect op de geschatte LTS regressierechte.

FIGUUR 7
Effect van uitschieters op de LTS schatter



De robuuste regressie schatters LMS en LTS hebben ook nadelen. Door te werken met medianen en afgeknotte sommen in de doelfunctie van (5) in plaats van met de volledige som der gekwadrateerde residuen, zullen deze schatters aan efficiëntie inboeten. Ze zijn met andere woorden minder precies dan de LS schatter wanneer er geen uitschieters zijn en de hypothese van normaliteit geldt. Daarom werden alternatieve robuuste schatters voorgesteld die efficiënter zijn dan LMS of LTS. Vaak kan men zulke schatters interpreteren als herwogen kleinste kwadraten schatters, waar de gewichten afhangen van de grootte van de residuen ten op zichte van een initiële LTS fit. Definities en implementaties van deze schatters vindt men bijvoorbeeld in Marazzi (1993).

IV. VOORBEELDEN

Lineaire regressie is een van de meest gebruikte statistische technieken. Het wordt op courante wijze gebruikt in de toegepaste economie. Toch wordt er weinig aandacht besteed aan het probleem van uitschieters wanneer men regressie toepast. Hieronder bespreken we kort twee voorbeelden van auteurs die in hun werk robuuste methoden gebruikten.

Het eerste voorbeeld is afkomstig van De Leval (2001) die actuariële pensioenplannen bestudeert. De berekeningen van deze pensioenplannen steunen op sterftetabellen, die op regelmatige basis aangepast worden. Het doel van de auteur was om waarden van toekomstige sterftetabellen te voorspellen, door de sterftetrend te modelleren. We beperken ons hier tot de kans dat een 30-jarige man zal sterven in het daaropvolgende jaar, voorgesteld als q_{30}. Deze kans verandert natuurlijk doorheen de tijd, en we definiëren q_{30}^t als de sterftekans in jaar t, waar het aantal jaren gemeten wordt vanaf het referentiejaar 1885. Volgend log-lineair model werd dan geponeerd

$$q_{30}^t = a\, b^t$$

voor $t = 1,\ldots,T$. Hier staat a voor het initiële overlijdensniveau in het referentiejaar en b staat voor het jaarlijks percentage verandering van de overlijdenskans. Op logaritmische schaal, en na toevoeging van een storingsterm bekomen we dan een eenvoudig lineaire model

$$y_t = \alpha + \beta\, t + \varepsilon_t$$

met

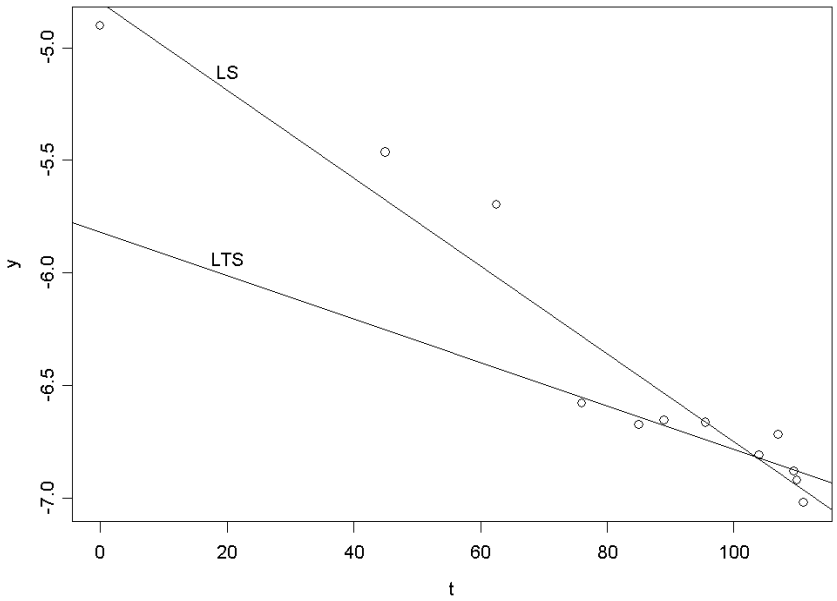
$$\alpha = \ln a,\ \beta = \ln b \text{ en } y_t = \ln q_{30}^t .$$

Sterftetabellen worden niet jaarlijks aangepast, maar worden constant gehouden gedurende bepaalde periodes. Voor twaalf verschillende periodes werden uit sterftetabellen de overlijdenskansen q_{30}^{t} bekomen, waar *t* overeenkomt met het midden van zo een periode. De laatste beschouwde periode was 1995-1997. We hebben dus geen jaarlijkse observaties, maar slechts 12 observaties die (hopelijk) representatief zijn voor de periodes. Figuur 8 toont de puntenwolk (t, y_t) met twee geschatte regressierechten gebaseerd op LS respectievelijk LTS-regressie. We zien dat de LTS een betere fit geeft voor de meer recente waarnemingen, dus deze met een grote waarden voor *t*. Voor de andere observaties, met kleine waarde van *t*, geeft LTS een minder goede fit, gezien LTS vooral de meerderheid van de data goed wil fitten. De robuuste methode zal de observaties, die het lineaire model minder goed volgen, een kleiner gewicht geven en in dit voorbeeld zijn dat de minst recente observaties. Merk op dat de eerste observatie een hefboompunt is. De reden waarom de oudste waarnemingen uitschieters zijn, is waarschijnlijk omdat er in de periode voor de tweede wereldoorlog een andere relatie tussen de sterftekansen en de tijd bestond. De robuuste analyse brengt aan het licht dat er twee structuren in de gegevens zijn, waar de klassieke analyse dit veel minder duidelijk zichtbaar maakt.

Stelt men zich even voor dat de meerderheid van de observaties zouden komen van de periode voor 1945. Dan zou de LTS schatter een goede fit geven voor de oudste observaties, en een minder goede fit voor de recentere. Maar, deze meer recente observaties zouden dan wel als een groep van uitschieters gedetecteerd worden. Indien men dan een voorspelling zou willen maken van een toekomstige sterftekans, is het duidelijk dat het model herschat moet worden op basis van enkel de meest recente observaties.

FIGUUR 8

Hierbij is $y = \log q_{30}^{t}$, *met* q_{30}^{t} *de kans op overlijden voor de 30-jarige Belgische mannen in jaar t, t.o.v. de tijd t sinds 1885, samen met een klassieke (LS) en robuuste (LTS) regressiefit*



Een ander voorbeeld van toepassing van robuuste regressie vinden we in het artikel van Knez en Ready (1997). Er blijkt enige controverse te zijn over het effect van de grootte van een bedrijf op het verwachtte rendement van zijn aandeel op de beurs. Gegevens kwamen van niet-financiële bedrijven, genoteerd op de New York Stock Exchange (NYSE), de American Stock Exchange (AMEX), en het Nasdaq-register van het Center for Research in Security Prices (CRSP) gedurende de periode van juli 1963 tot december 1990. Fama en French (1992) identificeren meerdere risicofaktoren die rendementsverschillen kunnen uitleggen, maar we beperken ons hier tot de factor "grootte", gemeten als het logaritme van de totale marktwaarde van de aandelen van het bedrijf. Terwijl Fama en French hun onderzoek baseren op kleinste kwadraten schattingen van het lineaire model, gebruiken Knez en Ready (1997) de robuuste LTS schatter.

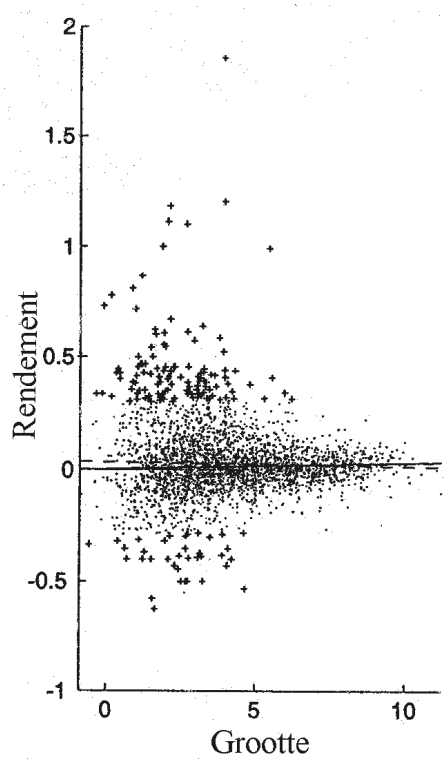
Voor de maand maart 1989, die representatief is voor andere maanden, stelt Figuur 9, genomen uit het artikel van Kenz en Ready (1997), de puntenwolk van de rendementen van de firma's in functie van hun groottes voor. De (lichtjes stijgende) volle rechte correspondeert met de LTS methode, terwijl de (lichtjes dalende) stippellijn correspondeert met de regressierechte gebaseerd op de LS methode. Gezien de hellingscoëfficiënten zeer klein zijn, is het verschil tussen de twee rechten klein, maar het blijkt significant verschillend te zijn en voor de meeste maanden voor te komen. De interpretatie is natuurlijk erg verschillend: hebben we een dalend of een stijgend rendement in functie van de grootte?

Er werd hier een 5% afgeknotte som van gekwadrateerde residuen als doelfunctie genomen voor de LTS-schatter. Er worden dus 5% van de observaties getrimd, en deze worden met een '+' symbool in het diagram van

Figuur 9 voorgesteld. Men merkt op dat deze observaties vooral voorkomen bij de eerder kleine bedrijven die een groot rendement halen. Er zijn hier dus een aantal, niet erg extreme, verticale uitschieters. De schatting bekomen met LS wordt naar boven vertekend door deze kleine bedrijven die grote positieve rendementen hebben, maar die toch maar minder dan 1% van de gegevens vertegenwoordigen. We verwijzen naar het artikel van Kenz en Ready (1997) voor meer detail.

FIGUUR 9

Rendement versus "grootte" van beursgenoteerde bedrijven,met een kleinste kwadraten (stippellijn) en een robuust LTS fit met α= 5% (volle lijn). De 5% "afgeknotte" observaties zijn aangeduid met een '+'



V. CONCLUSIES

In dit artikel hebben we getracht enkele basisbegrippen uit de theorie van de robuuste statistiek, zoals breekpunt en invloedsfunctie, op eenvoudige wijze uit te leggen. Verder werd een robuuste regressieschatter besproken die dan op twee economische voorbeelden werd geïllustreerd. We verwijzen naar

Zaman et al (2001) voor nog andere econometrische toepassingen van robuuste methoden.

Laat het duidelijk zijn dat we hier slechts kort hebben kunnen kennismaken met de theorie en praktijk van de robuuste statistiek. Basiswerken in het domein zijn Huber (1981) en Hampel et al (1986), die nog steeds verplichte literatuur zijn voor iedereen die in dit vakgebied werkt. Deze twee basiswerken zijn echter weinig op toepassingen gericht, en soms niet genoeg wiskundig rigoreus. Meer wiskundige werken, waar vooral aandacht is voor het limietgedrag van robuuste schatters en toetsen, vindt men in Rieder (1994) en Jureckova en Sen (1996). Een eerste boek in robuuste statistiek dat zich tot een breed publiek richtte, en zeker heeft bijgedragen tot een verdere doorbraak en verspreiding van het onderzoeksgebied, is Rousseeuw en Leroy (1987). Aan de hand van vele voorbeelden wordt hier op eenvoudige wijze de robuuste regressieproblematiek behandeld. Een ander toegankelijk werk is Staudte en Seather (1990), dat naast het regressiemodel ook nog veel aandacht besteed aan één- en twee-steekproef-problemen. Recentere werken zijn Wilcox (1997) en McKean en Hettmansperger (1998). Het eerste boek is erg praktijkgericht, vaak met een eigenzinnige keuze van de aangewende methoden, en het tweede focust op het gebruik van rang-methoden. Vermelden we nog het handboek van Madalla en Rao (1997), waarin meerdere robuuste statistische inferentie procedures behandeld worden.

Verder bestaat er ook een literatuur die procedures beschrijft om uitschieters en afwijkingen van een geponeerd regressiemodel te detecteren (bvb. Riani en Atkinson (2000), Chatterjee et al (2000), en Cook en Weisberg (1999)). We spreken hier van het domein van *Regression Diagnostics*. Merk op dat sommige van de voorgestelde procedures in deze literatuur enkel kunnen gebruikt worden indien er slecht één enkele uitschieter aanwezig is. Ze laten niet toe om het model te valideren indien er meerdere uitschieters aanwezig zijn. Het is hier niet de bedoeling om schatters te berekenen of toetsen uit te voeren. In het bekende boek van Draper en Smith (1998) over toegepaste regressieanalyse komt deze aanpak, samen met robuuste regressie, aan bod.

Een ander domein, gerelateerd tot robuustheid, is exploratieve gegevens-analyse. Hier worden grafieken en beschrijvende statistiek gebruikt om inzicht te krijgen in de structuur van de gegevens. Merk op dat we met beschrijvende statistiek geenszins het gebruik van eenvoudige schatters bedoelen, maar veeleer het berekenen van beschrijvende maten zonder expliciete referentie naar een statistisch model. Robuuste methodes, die de structuur van de meerderheid van de data zoekt en toelaat uitschieters te detecteren, is hier een natuurlijk hulpmiddel (zie Hoaglin et al (1982)).

Zoals reeds vermeld, is het mogelijk om robuuste regressie uit te voeren met bekende statistische softwarepakketten. Baanbrekend hierin is het pakket Splus, wat reeds vele jaren over een uitgebreide bibliotheek van robuuste procedures beschikt (Marazzi (1993)). Meer recent werden robuuste schatters in Stata en SAS opgenomen, we verwijzen hiervoor naar de web-documenten van Chen et al (2003) en SAS (2002).

Terwijl robuuste regressie goed bestudeerd is, zijn er nog vele andere statistische technieken waar de robuustheid nog verder voor ontwikkeld moet worden. Vooral in het domein van de multivariate statistiek, niet-lineaire veralgemeende regressie en tijdreeksmodellen is er nog veel werk te verrichten. Door de auteurs van dit artikel werden bijdragen geleverd in onder andere principaalcomponentenanalyse (Croux en Haesbroeck (2000)), factor modellen (Croux et al (2003)), logistische regressie (Croux en Haesbroeck (2003)), en discriminantanalyse (Croux en Dehon (2001), Croux en Joossens (2003)). Verder onderzoek is nog steeds lopend binnen onze onderzoeksgroep.

NOTEN

1. Een andere optie is om de hypothese van normaliteit op te geven en met een ander model, zoals bijvoorbeeld een Weibull verdeling, te werken.

REFERENTIES

Chatterjee, S.; Price, B. en Hadi, A., 2000, Regression Analysis by Example, (Wiley, New York).

Chen, X.; Ender, P.B.; Mitchell, M. en Wells, C, Regressions with STATA, (Stata Web Books), *http://www.ats.ucla.edu/stat/sas/webbooks/reg/*.

Cook, R.D. en Weisberg, S., 1999, Applied Regression Including Computing and Graphics, (Wiley, New York).

Croux, C. en Dehon, C. , 2001, Robust Linear Discriminant Analysis using S-estimators, *The Canadian Journal of Statistics* 29, 473-492.

Croux, C.; Filzmoser, P; Pison, G. en Rousseeuw, P.J., 2003, Fitting Multiplicative Models by Robust Alternating Egressions, *Statistics and Computing* 13, 23-36.

Croux, C. en Haesbroeck, G., 2000, Robust Principal Component Analysis based on Robust Estimators of the Covariance or Correlation Matrix, *Biometrika* 87, 603-618.

Croux, C., and Haesbroeck, G., 2002, Maxbias Curves of Location Estimators based on Subranges, *Journal of Nonparametric Statistics* 14, 295-306.

Croux, C. en Haesbroeck, G., 2003, Implementing the Bianco and Yohai estimator for Logistic Regression, *Computational Statistics and Data Analysis* 44, 273-295.

Croux, C. en Joossens, K., 2003, Influence of Observations on the Misclassification Probability in Quadratic Discriminant Analysis, Onderzoeksrapport 0359, (DTEW-KULeuven).

De Leval, D., 2001, Etude comptable et actuarielle des plans de pension: Confrontation des normes IAS/US GAAP et introduction de tables de mortalité prospectives, Mémoire de fin d'études sous la supervision de D. Justens, (Département de Gestion, Université de Liège).

Draper, N.R. en Smith, H., 1998, Applied Regression Analysis, (Wiley, New York).

Fama, E.F. en French, K.R., 1992, The Cross-Section of Expected Stocks Returns, *Journal of Finance* 47, 427-466.

Hampel, F.R., 1971, A General Qualitative Definition of Robustness, *Annals of Mathematical Statistics* 42, 1887-1896.

Hampel, F.R.; Ronchetti, E.M.; Rousseeuw, P.J. en Stahel, W.A., 1986, Robust Statistics: the Approach Based on Influence Functions, (Wiley, New York).

Hoaglin, D.A.; Mosteller, F. en Tukey, J.W., 1982, Understanding Robust and Exploratory Data Analysis, (Wiley, New York).

Huber, P.J., 1964, Robust Estimation of a Location Parameter, *The Annals of Mathematical Statistics* 35, 73-101.

Huber, P.J., 1981, Robust Statistics, (Wiley, New York).

Jureckova, J. en Sen, P.K., 1996, Robust Statistical Procedures: Asymptotics and Interrelations, (Wiley, New York).

KBC Bank en Verzekeringen, 2001, Economisch profiel van de Europese Unie, *Economische Financiële Berichten*, 8.
Knez, P.J. en Ready, M.J., 1997, On the Robustness of Size and Book-to-Market in Cross-Sectional Regressions, *Journal of Finance* 52, 1355-1382.
Maddala, G.S. en Rao, C.R., 1997, Handbook of Statistics 15: Robust Inference, (Elsevier, Amsterdam).
Marazzi, 1993, Algorithms, Routines, and S-Functions for Robust Statistics, (Champan and Hall, New York).
McKean, J.W. en Hettmansperger, T.P., 1998, Robust Nonparametric Statistical Methods, (Arnold, London).
Riani, M. en Atkison, A., 2000, Robust Diagnostic Regression Analysis, (Spinger, Berlin).
Rieder, H., 1994, Robust Asymptotic Statistics, (Springer, Berlin).
Rousseeuw, P.J., 1984, Least Median of Squares Regression, *Journal of the American Statistical Association* 79, 871-880.
Rousseeuw, P.J. en Leroy, A.M., 1987, Robust Regression and Outlier Detection, (Wiley, New York).
SAS OnlineDoc., 2002, IML: Robust Regression, (Sas Institute, Cary, NC), *http://v8doc.sas.com/sashtml*.
Staudte, R.G. en Seather, S.J., 1990, Robust Estimation and Testing, (Wiley, New York).
Wilcox, R., 1997, Introduction to Robust Estimation and Hypothesis Testing, (Academic Press, San Diego).
Zaman, A.; Rousseeuw, P.J. en Orhan, M., 2001, Econometric Applications of High-Breakdown Robust Regression Techniques, *Economics Letters* 71, 1-8.